



Recherche automatique d'hétéro-répétitions dans un dialogue oral spontané

Brigitte Bigi, Roxane Bertrand, Mathilde Guardiola

► To cite this version:

Brigitte Bigi, Roxane Bertrand, Mathilde Guardiola. Recherche automatique d'hétéro-répétitions dans un dialogue oral spontané. XVIIIèmes Journées d'étude sur la Parole, May 2010, Mons, Belgique. pp.1-4. hal-00541976

HAL Id: hal-00541976

<https://hal.science/hal-00541976>

Submitted on 1 Dec 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Recherche automatique d'hétéro-répétitions dans un dialogue oral spontané

Brigitte Bigi, Roxane Bertrand, Mathilde Guardiola

Laboratoire Parole & Langage, CNRS & Aix-Marseille Universités
5, avenue Pasteur, BP 80975, 13604 Aix en Provence, France
Mél : {brigitte.bigi, roxane.bertrand, mathilde.guardiola}@lpl-aix.fr

ABSTRACT

Other-repetitions are a device involving the reproduction by a speaker of what another speaker has just said. This paper proposes a solution to automatically detect other-repetitions in French conversational dialogue. A first step of the proposed system consists in finding all possible other-repetitions in the dialogue. A second step is used to select other-repetitions which need to be kept by combining rules with speaker statistics. This automatic detection, evaluated on a one hour dialogue, shows good results according to the expected objectives : recall is 1, and precision is about 80%.

Keywords: other-repetitions, automatic detection, conversational dialogue, French.

1. INTRODUCTION

Ce papier présente une méthode d'identification automatique des hétéro-répétitions ou répétitions diaphoniques [11], c'est-à-dire la reproduction par un locuteur 2 d'un énoncé ou d'une partie d'énoncé préalablement produit par un locuteur 1. La majorité des travaux relatifs aux répétitions concerne essentiellement les auto-répétitions [5], lesquelles sont fréquemment associées aux phénomènes de disfluences du discours (voir entre autres [6, 7]). Les hétéro-répétitions restent quant à elles encore peu explorées. A notre connaissance, il n'existe pas de travaux relatifs à leur identification automatique. Le but d'un tel outil est de proposer un nombre réduit de segments du dialogue susceptibles d'être des répétitions, de sorte que seuls ces derniers soient examinés manuellement. Disposer d'un tel outil se révèle donc une aide précieuse pour l'annotation de corpus de dialogues de plus en plus importants en taille. Il permet en outre de systématiser et de comparer des données nombreuses extraites de corpus variés de dialogues (débat, entretiens, etc.).

L'intérêt porté aux hétéro-répétitions s'inscrit dans le cadre plus large d'un projet sur l'imitation dans la parole (projet ANR SPIM¹). Les répétitions, phénomène universel et particulièrement crucial dans la communication humaine, constituent en ce sens l'un des procédés les plus explicites pour explorer cette question, en témoignent les nombreux travaux menés en acquisition du langage notamment (pour une revue, voir [5]) et le rôle indiscutable de l'imitation pour améliorer les compétences linguistiques du petit enfant. D'un point de vue pragmatique, des résultats montrant que les répétitions joueraient davan-

tage encore un rôle dans l'acquisition d'une compétence communicative (versus une compétence linguistique) chez l'enfant, ont toutefois conduit les auteurs à distinguer répétition et imitation [4]. Au delà de cette dernière, les répétitions serviraient donc d'autres objectifs communicatifs. Nous nous intéressons précisément aux hétéro-répétitions pour les différentes fonctions pragmatiques qu'elles revêtent dans l'interaction. Parmi d'autres, les répétitions constituent des procédés discursifs grâce auxquels les locuteurs convergent, s'alignent (ou pas) pour répondre au principe de coopération qui régit toute conversation. En effet, la conversation se définit notamment par un très haut degré de coopération qui n'exclut cependant pas des moments de tension, voire de compétition. Mais les participants à une conversation n'ont de cesse de collaborer pour construire ensemble à la fois le sens mais aussi la relation qui définit l'interaction en cours. Cette coopération est visible au travers notamment de procédés discursifs spécifiques, et renvoie à la question centrale du ménagement des faces en présence qui conditionne l'activité de régulation des discours.

Les backchannels - signaux multimodaux tels que *mh*, *ouais*, *hochement de tête*, *sourire*, etc. émis par l'interlocuteur en dialogue - participent à cette régulation en informant sur le processus d'écoute et de compréhension des discours [8]. Les hétéro-répétitions jouent dans ce cadre un rôle crucial. Certains auteurs, à l'image de Laforest [9], les considèrent comme une catégorie complexe² de backchannels en raison notamment des fonctions pragmatiques qui leur sont associées. Il existe diverses classifications fonctionnelles des signaux backchannels, dans lesquelles les deux classes de « continuer » (rôle d'accusé-réception) et d'« assessment » (rôle de prise de position) demeurent consensuelles [12]. Plus exhaustive et détaillée, la classification de Maynard [10] ajoute une fonction de compréhension, une fonction de support/adhésion au discours, et une fonction de demande de confirmation. La typologie établie par Perrin et al., dans [11] confirme un fonctionnement très similaire des hétéro-répétitions en français québécois à travers les 4 fonctions suivantes :

- « a taking into account function », signalant que l'interlocuteur a correctement entendu et interprété le discours précédent,
- « a confirmation request function », signalant un problème dans le discours,

¹ Imitation in speech : From sensori-motor integration to the dynamics of conversational interaction. <http://lpl-aix.fr/projet/169>

² L'auteur oppose les régulateurs simples (*mh*, *ouais*, etc) qui ne sont pas considérés comme de réels tours de parole (au sens d'apport informatif) aux régulateurs complexes qui renvoient non seulement aux répétitions mais aussi aux reformulations, compléments, et métaquestions dont le statut de non tour s'avère plus délicat à établir.

- « a positive reply function », qui signale l'accord de l'interlocuteur avec le discours précédent,
 - « a negative reply function », signalant le désaccord.
- Dans une étude pragmatique et prosodique, [3] ont pu mettre en évidence une fonction spécifique des hétéro-répétitions qui joueraient un rôle fondamental dans la construction collective de séquences discursives humoristiques. Nous avons pour but d'étudier les hétéro-répétitions pour mettre ainsi à jour de nouvelles fonctions pragmatiques susceptibles de mieux caractériser certains moments de convergence dans l'interaction.

Dans la section suivante, nous présentons les critères formels de définition des répétitions sur la base desquels l'outil a été conçu. La prochaine section décrit la méthode implémentée pour la recherche automatique. Nous présentons enfin brièvement le corpus pour lequel l'outil a été développé, puis une évaluation de ce dernier, sur l'un des dialogues de ce corpus, et des exemples illustrant les catégories fonctionnelles proposées dans [11].

2. CRITÈRES FORMELS DE DÉFINITION DES RÉPÉTITIONS : ÉTAPE PRÉALABLE

Une observation préalable d'un dialogue a permis d'identifier les critères formels délimitaires des hétéro-répétitions sur lesquels reposera l'implémentation de l'outil de recherche automatique. L'identification a été faite sur une base lexicale (répétition de mots), le long d'un continuum allant des répétitions lexicales à l'identique (répétition verbatim ou écho strict) aux répétitions partielles ou écho partiel (modification d'un item, changement de pronom, ajout d'une particule discursive, etc.). L'outil automatique prend ainsi en considération le fait qu'une répétition est en écho strict ou non (section 3.2). Les phénomènes de reformulation, parfois associés aux phénomènes de répétition, ont en revanche été éliminés. Un autre critère d'identification concerne la taille de la répétition. Nous nous sommes demandé dans quelle mesure on peut considérer que la reprise d'un seul mot constitue une réelle répétition. L'une des options a été de considérer comme relevant d'une répétition un mot qui s'avère relativement rare dans le dialogue. Nous verrons que l'outil de recherche automatique a permis d'affiner cette notion puisqu'elle a impliqué la prise en compte de ce critère de fréquence de mot chez l'un et l'autre des locuteurs (section 3.2). Enfin cette exploration manuelle préalable nous a permis de mettre à jour un autre critère délimitaire des répétitions souvent omis lorsque l'on définit les répétitions comme étant la reproduction presque immédiate de l'énoncé précédent. Il s'agit de l'empan de discours sur lequel la répétition doit être recherchée. Nos premières observations ont permis de mettre à jour des répétitions (verbatim notamment), relativement longues, sur un empan assez important du discours, à savoir sur plusieurs tours de parole. L'outil automatique intègre ce critère fondamental (section 3.1).

3. RECHERCHE AUTOMATIQUE DES HÉTÉRO-RÉPÉTITIONS

Les travaux présentés dans [1] concernent la détection et la correction automatique des auto-répétitions dans un contexte de dialogue homme-machine. Une correspondance lexicale systématique permet de détecter un ensemble initial de répétitions candidates. Celles-ci sont ensuite analysées avec des informations de niveaux syntaxique, sé-

mantique et acoustique afin de distinguer les véritables répétitions des fausses détections. Si la démarche est intéressante, la différence majeure avec le travail que nous présentons réside dans le fait que les répétitions recherchées dans [1] sont intra-locuteurs, par exemple « show me *flights* daily *flights* to boston ». Nous proposons une recherche des énoncés « source » (énoncés sur lesquels s'ancrent les énoncés répétés), filtrés selon des règles incluant des statistiques, qui reprennent et systématisent les critères identifiés manuellement.

3.1. Recherche systématique

Pour effectuer la recherche systématique des répétitions, nous nous appuyons sur la segmentation du dialogue en IPU - Inter-Pausal Units, blocs de parole bornés par des pauses silencieuses de 200 ms, ainsi que sur leur alignement sur le signal qui donne leur localisation temporelle. La recherche des hétéro-répétitions consiste à mettre en correspondance :

- les mots d'un locuteur d'une IPU donnée, qui correspond à une localisation temporelle dans le signal ;
- avec les mots de l'autre locuteur, sur une IPU de localisation temporelle proche et jusqu'à un nombre N d'IPU suivantes.

En théorie, toutes les IPU qui ne sont pas sélectionnées par cette recherche systématique ne peuvent pas contenir de répétitions puisque les deux locuteurs n'emploient aucun vocabulaire commun. Or, nécessairement, en dialogue, un énoncé produit à la première personne par exemple sera répété par l'interlocuteur à la seconde, ce qui en fait des répétitions authentiques mais qui répondent aux contraintes du dialogue. Ainsi, dans le but d'obtenir un taux de faux rejets minimum, une lemmatisation permet de capter des répétitions qui ne l'auraient pas été par l'utilisation des mots orthographiés. Elle permet en effet de mettre en correspondance des verbes conjugués différemment par les deux locuteurs, des mots au singulier ou pluriel, tous les pronoms sujets sont ramenés à la forme *il*, etc. La lemmatisation permet d'une part de trouver des répétitions supplémentaires, d'autre part de renvoyer des segments plus longs, comme dans l'exemple ci-après. L'utilisation des mots permettrait de renvoyer une répétition de 2 mots :

source : non j' ai pas voulu non
écho : tu avais pas voulu

La lemmatisation permet de renvoyer une suite de 4 lemmes :

source : non il avoir pas vouloir non
écho : il avoir pas vouloir

Dans la suite de ce document, nous considérons non pas le lexème en tant que tel mais le lemme ; le « mot » fera référence à la lemmatisation de celui-ci.

Avec un empan temporel suffisamment large, cette première étape produit l'ensemble des répétitions possibles. L'intérêt d'un outil automatique réside aussi en sa capacité de ré-itérer les exécutions afin de déterminer les valeurs les plus appropriées. Pour notre corpus, après essai de différentes valeurs, nous avons choisi de chercher dans les 3 IPU qui suivent l'émission. Cependant, les IPU sélectionnées contiennent de nombreuses correspondances qui ne sont pas des répétitions (utilisation par les deux locuteurs du même déterminant, d'un même article par exemple). De fait, la seconde étape filtre les répétitions repérées.

3.2. Filtrage des répétitions

La sélection repose sur l'utilisation de deux règles :

Règle 1 Une répétition candidate est acceptée dans la mesure où elle contient au moins un mot pertinent du point de vue du locuteur qui répète.

Règle 2 Une répétition longue (au moins K mots) composée seulement de mots non pertinents sera acceptée, dès lors qu'elle est répétée telle quelle (en écho strict). Par expérience, la valeur optimale de K , pour notre corpus, est 3.

La règle 1 nécessite de préciser la notion de mot pertinent d'un locuteur. Introduire ce critère pour déterminer si une répétition doit être sélectionnée est essentiel. Sans ce critère, beaucoup de répétitions de mots considérés comme « usuels » seraient conservées. Il s'agit donc d'établir, soit une liste de mots pertinents, soit au contraire, une liste de mots non-pertinents. Il serait possible d'utiliser un lexique *a priori* des mots outils de la langue, par exemple, comme cela est déjà fait pour l'écrit. Cependant, nous avons constaté que certains mots peuvent être pertinents pour un locuteur, sans l'être pour un autre. De même, certains mots généralement considérés comme mots outils s'avèrent pertinents pour un locuteur car très peu employés. Ceci sera illustré dans la section relative à l'évaluation. Nous avons alors opté pour une sélection dynamique des mots pertinents : une liste est établie pour chaque locuteur, dans chaque dialogue.

On note $N_l(w)$, le nombre d'occurrences du mot w prononcé par le locuteur l , et $|V_l|$, la taille du vocabulaire (nombre de mots différents) du locuteur l . On définit enfin $P_l(w)$, la probabilité du mot w pour le locuteur l par :

$$P_l(w) = \frac{N_l(w)}{\sum_i |V_l| N_l(w_i)}$$

On dira qu'un mot w est pertinent pour un locuteur l , si :

$$P_l(w) \leq \frac{1}{\alpha \times |V_l|}$$

Le coefficient α pourra être déterminé de façon empirique selon le type de corpus utilisé. Dans nos expériences, nous avons utilisé $\alpha = \frac{1}{2}$. D'un certain point de vue, on peut considérer qu'un mot est pertinent s'il n'est pas très fréquent pour ce locuteur, au regard de la richesse de son vocabulaire.

L'utilisation conjointe de règles simples et de statistiques directement déduites des IPU de chaque locuteur constitue ainsi une solution simple pour filtrer automatiquement les hétéro-répétitions.

4. CONTEXTE APPLICATIF

4.1. Description du corpus

Nos données sont extraites du CID, corpus conversationnel composé de 8 dialogues. Chacun de ces dialogues dure une heure et implique deux participants. Un travail important autour de ce corpus a été l'élaboration d'un schéma d'annotation pour l'ensemble des niveaux de l'analyse linguistique (morpho-syntaxique, syntaxique, phonétique, prosodique, discursive) et des différentes modalités impliquées dans le dialogue (audio et mimo-gestuel) [2]. Nous nous limiterons à présenter ici le premier niveau d'annotation qu'est la transcription étant donné que l'outil d'identification des hétéro-répétitions se fonde exclusivement

sur ce niveau. Sans entrer dans le détail, soulignons que cette transcription a été effectuée pour traiter des données multi-niveaux. Nous avons donc opté pour une transcription dite orthographique enrichie (TOE) afin de garantir à la fois un meilleur rendement de l'alignement phonétique nécessaire à l'analyse des modules phonétique et prosodique, et le meilleur rendement pour les modules morpho-syntaxique et syntaxique qui ne peuvent être faits que sur des transcriptions orthographiques standard (utilisation de lexiques, etc.). Le CID a été automatiquement segmenté en IPU ; chacune des 12950 IPU est alors alignée phonétiquement indépendamment, ce qui évite de propager d'éventuelles erreurs au delà d'une IPU. Les transcriptions, effectuées sur les 8 heures, se sont déroulées en deux phases : la première a consisté en la transcription et la correction des dialogues par 2 annotateurs, la seconde a été effectuée par un expert qui a corrigé une nouvelle fois l'ensemble des transcriptions.

L'évaluation des performances de l'outil automatique est réalisée manuellement, elle porte sur l'un des dialogues du CID. Néanmoins, ce dialogue contient un nombre suffisant d'hétéro-répétitions pour que cette évaluation soit pertinente. Les locuteurs sont mentionnés par AB et CM. La table 1 présente quelques valeurs textométriques pour ces locuteurs. On constate que le vocabulaire de AB est plus riche que celui de CM. En effet, même si AB intervient moins dans le dialogue (nombre d'occurrences inférieur) que CM, son vocabulaire est plus large et le nombre d'hapax³ est également plus élevé.

TAB. 1: Description du corpus en terme de vocabulaire

Loc.	Vocab.	Occ.	Hapax	Occ. max	Mot max
AB	1183	6619	671	184	on
CM	1066	7681	561	501	ouais

Selon la définition de la pertinence, proposée en section 3.2, un mot est pertinent s'il est apparu moins de 16 fois pour AB, 20 fois pour CM. Ainsi, le mot *petit*, présent 21 fois pour AB, mais 8 fois pour CM, est pertinent seulement pour CM. Inversement, le mot *voilà* est prononcé 8 fois par AB et 27 fois par CM ; il est donc pertinent seulement pour AB. De même, le mot *vachement* est apparu 2 fois pour AB, 20 fois pour CM. Des mots comme *le*, *de*, *pouvoir*, *truc* ne sont pas pertinents pour les deux locuteurs.

4.2. Evaluation du système automatique

La valeur de rappel a été obtenue par un expert qui a examiné manuellement toutes les répétitions trouvées dans l'empan temporel choisi. Comme nous le souhaitions, notre système a obtenu une valeur de rappel égale à 1 (aucune répétition omise). La valeur de précision a également été évaluée sur le dialogue AB-CM, grâce à l'analyse manuelle de deux experts sur les répétitions filtrées.

La table 2 indique le nombre de répétitions trouvées puis filtrées par le système automatique ainsi que le nombre d'entre-elles qui ont été validées manuellement par les experts. Dans le cas AB émet - CM répète, la précision du

³ mots présents une seule fois

système automatique est de 80,17 %. Dans le cas AB répétition - CM émet, elle est de 74,79 %. Il est important de noter que les IPU présentent de nombreux chevauchements de localisations, et dans ce cas, la même répétition est répétée dans les 2 sens .

TAB. 2: Répétitions automatiques et évaluation

Locuteurs	Nombre de répétitions		
	trouvées	filtrées	validées
AB émet - CM répétition	860	116	93
AB répétition - CM émet	1251	119	89

L'évaluation de cet outil est donc satisfaisante au regard des objectifs pour lesquels il a été conçu. Ses performances ne sont pas remises en question par les experts. Les différences sont en effet liées à des critères non automatisables. Par exemple, une répétition est pertinente si elle a un caractère ostensif, c'est-à-dire qu'elle est le fruit d'une intention du locuteur qui manifeste par cette répétition son intention de citer les mots en question [11]. En outre, un tel critère suppose que la répétition comporte une réaction de l'interlocuteur sur ce qu'il cite.

4.3. Illustrations

Cette section présente quelques exemples de répétitions que l'outil a trouvé dans le dialogue AB-CM. Le premier exemple montre une répétition proposée par l'outil de détection automatique, car oui c' était est une répétition de 3 mots non pertinents en écho strict (application de la règle 2). Celle-ci n'a cependant pas été validée par l'expert :

CM oui c' était un insolite donc un peu malaise quand même quoi un peu désagréable quoi quand même ouais

AB oui c' était pas on n' était pas tristes mais on était on était coincés obligés alors qu' on était en vacances et

CM ouais ouais ouais ouais ouais

CM coincée ouais

Par ailleurs, cet exemple montre à nouveau l'intérêt d'utiliser des lemmes, car ils permettent de repérer la répétition du verbe coïncider. L'exemple ci-après illustre également l'intérêt d'utiliser des lemmes :

CM et il contrôlait pas

AB il a pas contrôlé

L'exemple suivant montre un cas où la répétition est formulée sur deux IPU.

CM c' était La Rochelle

AB non c' était à Poitiers

CM c' était à

CM à Poitiers tu as été à Poitiers ouais

CM ouais

Enfin, si la grande majorité des répétitions détectées puis validées par les experts concernent moins de 6 mots, nous en avons également identifiées de très longues comme celle reproduite ci-dessous qui présente un écho strict de 11 items :

CM et y a de la neige qui était rentrée dans la chambre et

AB et y a de la neige qui était rentrée dans la fenêtre oui un peu

5. CONCLUSION

Cet article concerne la recherche des hétéro-répétitions en conversation. La démarche qui nous a amené à les identifier ainsi que l'outil automatique qui permet de les repérer ont été présentés. De nombreux exemples, ainsi qu'une évaluation de la méthode, permettent de valider l'approche automatique ainsi que la pertinence de son utilisation au sein d'un dialogue oral spontané. A court terme, cet outil permettra de constituer une base de données conséquente d'hétéro-répétitions puisque nous souhaitons l'utiliser sur plusieurs types de corpus (comme les débats, les entretiens, les dialogues orientés-tâche, etc.) en vue de comparer les répétitions en termes formels et fonctionnels.

RÉFÉRENCES

- [1] J. Bear, J. Dowding, and E. Shriberg. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *30th annual meeting on Association for Computational Linguistics*, pages 56–63, Newark, Delaware, 1992.
- [2] R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, and S. Rauzy. Le cid - corpus of interactional data. *Traitement Automatique des Langues*, 49(3) :105–134, 2008.
- [3] R. Bertrand and B. Priego-Valverde. Does prosody play a specific role in conversational humor? *Pragmatics and Cognition*, 18(2), 2010.
- [4] M-W. Casby. A pragmatic perspective of repetition in child language. *Journal of Psycholinguistic Research*, 15(2) :127–140, 1986.
- [5] H. Chiung-chih. Other-repetition in mandarin child language : A discourse pragmatic perspective. *Journal of Pragmatics*, doi :10.1016/j.pragma.2009.08.005, 2009.
- [6] J. Cole, M. Hasegawa-Johnson, C. Shih, H. Kim, E-K. Lee, H-Y. Lu, Y. Mo, and T-J. Yoon. Prosodic parallelism as a cue to repetition disfluency. In *Disfluency in Spontaneous Speech Workshop*, Aix-en-Provence, France, 2005.
- [7] T-S. Curl, J. Local, and G. Walker. Repetition and the prosody-pragmatics interface. *Journal of Pragmatics*, 38 :1721–1751, 2006.
- [8] J-E. FoxTree. Listening in on monologues and dialogues. *Discourse Processes*, 27(1) :35–53, 1999.
- [9] M. Laforest. Le back-channel en situation d'entrevue. *Québec : CIRAL/Recherches sociolinguistiques*, 2, 1992.
- [10] S. Maynard. Japanese conversation : Self-contextualization through structure and interactional management. Ablex, Norwood, NJ, 1989.
- [11] L. Perrin, D. Deshaies, and C. Paradis. Pragmatic functions of local diaphonic repetitions in conversation. *Journal of Pragmatics*, 35 :1843–1860, 2003.
- [12] E-A. Schegloff. Discourse as an interactional achievement : Some uses of uh huh and other things that come between sentences. *Analyzing discourse : Text and talk*, pages 71–93, 1982.